Cross-domain Paraphrasing For Improving Language Modelling Using Out-of-domain Data

X. Liu, M. J. F. Gales & P. C. Woodland

Cambridge University Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {*xl207,mjfg,pcw*}*@eng.cam.ac.uk*

Abstract

In natural languages the variability in the underlying linguistic generation rules significantly alters the observed surface word sequence they create, and thus introduces a mismatch against other data generated via alternative realizations associated with, for example, a different domain. Hence, direct modelling of out-of-domain data can result in poor generalization to the in-domain data of interest. To handle this problem, this paper investigated using cross-domain paraphrastic language models to improve in-domain language modelling (LM) using out-of-domain data. Phrase level paraphrase models learnt from each domain were used to generate paraphrase variants for the data of other domains. These were used to both improve the context coverage of in-domain data, and reduce the domain mismatch of the out-of-domain data. Significant error rate reduction of 0.6% absolute was obtained on a state-of-the-art conversational telephone speech recognition task using a cross-domain paraphrastic multi-level LM trained on a billion words of mixed conversational and broadcast news data. Consistent improvements on the in-domain data context coverage were also obtained.

Index Terms: language model, paraphrase, speech recognition

1. Introduction

In natural languages multiple word sequences can represent the same underlying meaning. The mapping from the meaning to surface form involves a natural language generation process and is often one-to-many. The resulting surface realizations are paraphrastic to one other, but use different linguistic rules in generation. They represent different domains, styles or other speaker specific characteristics. The variability in these rules can significantly alter the observed surface word sequence, and thus introduce a mismatch against other data generated via alternative realizations associated with, for example, a target domain of interest. Hence, direct modelling of the observed surface word sequence found in out-of-domain data, can result in poor generalization to in-domain data, for example, when using n-gram language models (LM). As the diversity among surface forms found in different domains increases, convectional techniques using model or data combination become less effective [23, 25, 8, 9, 10, 24, 7, 14].

To handle this problem, it is possible to structurally exploit the generation rules associated with domain independent and dependent characteristics of the training data. Two approaches may be considered. First, the out-of-domain data, often available in large quantities, can be used to learn a rich set of domain independent linguistic generation rules that represent, for example, the paraphrastic relationships between different words, phrases or sentences [1, 18, 12, 22, 16]. Applying the resulting domain independent paraphrase models to the in-domain training data, often in limited quantities, can produce rich paraphrase variants to improve the in-domain LM context coverage. Second, the in-domain data can be viewed as a "degraded" form of the out-of-domain data generated via alternative surface realizations that represent, for example, the disfluency and informal style found in conversational speech. Hence, it is also possible to applying these paraphrastic mappings learnt from the in-domain data to the out-of-domain data to generate their associated "in-domain like" paraphrases.

Along these lines, this paper investigated using cross-domain paraphrastic language models to improve language modelling for conversational telephone speech (CTS) using out-of-domain broadcast news (BN) data. In order to increase both the context coverage of the conversational data, and reduce the domain mismatch of the broadcast news data, phrase level paraphrase models separately learnt from the data of each domain is used to generate paraphrase variants for the data of the other domain. These variants are then used in paraphrastic LM training. The rest of the paper is organized as follows. Paraphrastic language models are reviewed in section 2. The paraphrase extraction and lattice generation schemes are presented in section 3. Two cross domain paraphrase generation approaches are proposed in section 4. In section 5 a range of paraphrastic LMs are evaluated on a state-of-the-art conversational telephone speech transcription task. Section 6 is the conclusion and possible future work.

2. Paraphrastic Language Models

In order to capture the paraphrastic relationship between longer span syntactic structures, a more general form of modelling is preferred. The particular type of LMs considered in this paper can flexibly model paraphrase mapping at the word, phrase and sentence level. As LM probabilities are estimated in the paraphrased domain, they are referred to as *paraphrastic language models* (PLM) [16, 17]. For a L word long word sequence $\mathcal{W} =< w_1, w_2, ..., w_i, ..., w_L >$ in the training data, rather than maximizing the surface word sequence's log-probability $\ln P(\mathcal{W})$ as for conventional LMs, the marginal probability over all paraphrase variant sequences is maximized,

$$\mathcal{F}(\mathcal{W}) = \ln \left(\sum_{\boldsymbol{\psi}, \boldsymbol{\psi}', \mathcal{W}'} P(\mathcal{W}|\boldsymbol{\psi}) P(\boldsymbol{\psi}|\boldsymbol{\psi}') P(\boldsymbol{\psi}'|\mathcal{W}') P_{\text{PLM}}(\mathcal{W}') \right) \quad (1)$$

where

- $P_{\text{PLM}}(\mathcal{W}')$ is paraphrastic LM probability to be estimated.
- $P(\boldsymbol{\psi}'|\mathcal{W}')$ is a word to phrase segmentation model assigning the probability of a phrase level segmentation,

The research leading to these results was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology) and DARPA under the Broad Operational Language Translation (BOLT) program.

$\boldsymbol{\psi}'$, given a paraphrase word sequence $\mathcal{W}'$;

- $P(\boldsymbol{\psi}|\boldsymbol{\psi}')$ is a phrase to phrase paraphrase model computing the probability of a phrase sequence $\boldsymbol{\psi}$ being paraphrastic to another $\boldsymbol{\psi}'$;
- $P(\mathcal{W}|\boldsymbol{\psi})$ is a phrase to word segmentation model that converts a phrase sequence $\boldsymbol{\psi}$ to a word sequence $\mathcal{W}$, and by definition is a deterministic, one-to-one mapping, thus considered non-informative.

It can be shown that the sufficient statistics for a maximum likelihood (ML) estimation of $P_{\mathsf{PLM}}(\mathcal{W}')$ are accumulated along each paraphrase word sequence and weighted by its posterior probability. For a particular n-gram predicting word w_i following history h_i, the associated statistics $C(h_i, w_i)$ are

$$C(h_i, w_i) \;=\; \sum_{\mathcal{W}'} P(\mathcal{W}'|\mathcal{W}) C_{\mathcal{W}'}(h_i, w_i) \qquad (2)$$

where $C_{\mathcal{W}'}(h_i, w_i)$ is the count of subsequence $<h_i, w_i>$ occurring in paraphrase variant $\mathcal{W}'$. During word to phrase segmentation, ambiguity can occur. If there is no clear reason to favor one phrase segmentation over another, $P(\boldsymbol{\psi}'|\mathcal{W}')$ may be treated as non-informative, as is considered in this work.

As sufficient statistics are discounted and re-distributed to alternative expressions of the same word sequence, paraphrastic LMs are expected to have a richer context coverage and broader distribution, but at the same time potentially increased modelling confusion than conventional LMs trained on the surface word sequence. One approach to balance the specific, but poorer coverage word-based N-gram LMs with a more generic LM is to linearly interpolate the LM probabilities. This is commonly used with class-based LMs [20] and is used in this paper with paraphrastic LMs. Let $P(\tilde{w}|\tilde{h})$ denote the interpolated LM probability for any in-vocabulary word $\tilde{w}$ following an arbitrary history $\tilde{h}$, this is given by

$$P(\tilde{w}|\tilde{h}) \;=\; \lambda_{\mathsf{NG}} P_{\mathsf{NG}}(\tilde{w}|\tilde{h}) + \lambda_{\mathsf{PLM}} P_{\mathsf{PLM}}(\tilde{w}|\tilde{h}) \qquad (3)$$

where λ_{NG} and λ_{PLM} are the interpolation weights assigned to the conventional LM distribution $P_{\mathsf{NG}}(\cdot)$ and the paraphrastic LM $P_{\mathsf{PLM}}(\cdot)$. They can be optimized on some held-out data.

In order to increase the context span for paraphrastic LMs, a phrase level paraphrastic LM can also be trained. This can be obtained by optimizing a simplified form of criterion given in equation (1), where the word to phrase segmentation model $P(\boldsymbol{\psi}'|\mathcal{W}')$ is dropped, thus the sufficient statistics in equation (2) accumulated on phrase level instead. In order to incorporate richer linguistic constraints, it is possible to train and log-linearly combine LMs that model different units, for example, words and phrases. LMs built at word and phrase level are log-linearly combined to yield a multi-level LM to further improve discrimination [13, 14, 15]. This requires word level lattices to be first converted to phrase level lattices before the log-linear combination is performed. The log-linear interpolation weights were set as 0.6 and 0.4 for word and phrase level LMs, and kept fixed for all experiments of this paper.

3. Paraphrase Learning and Generation

As discussed in sections 1 and 2, a phrase level paraphrase model is used in paraphrastic LMs. In order to obtain sufficient phrase coverage, an appropriate technique to learn a large number of paraphrase phrase pairs is required. Since it is impractical to obtain expert semantic labelling at the phrase level, statistical paraphrase extraction schemes are needed [1, 18]. Hence, techniques that perform paraphrase pair extraction from standard text data [12, 22] are used. These are motivated by the *distributional similarity* theory [6], which postulates that phrase pairs often sharing the same left and right contexts are likely to be paraphrases to each other. As standard text data in large amounts can be used, wide phrase coverage can be obtained. Due to this advantage, the following n-gram paraphrase induction algorithm [16] is used to estimate the paraphrase model. The minimum and maximum phrase length are set as $L_{\min} = 1$ and $L_{\max} = 4$, and the left and right context length set as $L_N = 3$ and kept fixed for all experiments in this paper.

initialize phrase pair list $V = \{\}$;
initialize n-gram subsequence list $U = \{\}$;
for every sentence in training data **do**
    extract all variable length n-gram phrases together with their fixed length left and right contexts $\{c_l, v, c_r\}$
end for
for every n-gram phrase pair $<v \rightarrow v'>$ **do**
    compute the co-occurrence counts $C(v \rightarrow v')$ of them sharing the same left and right contexts c_l, c_r
end for
for every n-gram phrase pair $<v \rightarrow v'>$ **do**
    estimate paraphrase probabilities $p(v'|v) = \frac{C(v \rightarrow v')}{\sum_{\bar{v}} C(v \rightarrow \bar{v})}$
end for

The above algorithm can be extended to incorporate additional useful information, for example, syntactic constraints, in order to improve the grammaticality of paraphrase variants. In common with other paraphrase induction methods, the above scheme can also produce phrase pairs that are non-paraphrastic, for example, producing antonyms. However, this is of less concern for language modelling, for which improving context coverage is the prime aim.

In order to train paraphrastic LMs, multiple paraphrase variants are required to compute the sufficient statistics given in equation (2).These variants can be efficiently generated using a weighted finite state transducers (WFST) [19] based decoding approach [16], rather than designing special purpose decoding tools. The statistics required for paraphrastic LM estimation are then accumulated from the paraphrase lattices via a forward-backward pass. In order to improve phrase coverage, expert semantic labelling provided by resources, such as WordNet [5], can also be used to generate paraphrases [16, 17].

4. In-domain and Cross-domain Paraphrase Generation

As discussed above, the paraphrase variants used in PLM training were generated using a paraphrase model and some observed surface word sequences to be paraphrased. In previous research, both were obtained from the same data source [16, 17], and thus a common domain. For an example, using a paraphrase model trained on 545 million words of conversational data, for an in-domain sentence "*And I generally prefer*", the following paraphrase variants are among those generated:

Original sentence: *And I generally prefer*
In-domain Paraphrases:

And I really like	*I mean I would like*
I guess I generally like	*You know I just want*
So I appreciate	*I think I need*

'Cause I love	*Well I prefer*
Um I wish	*Yeah I just prefer*

Similarly for a sentence in the broadcast news domain, "*Economy is a big problem for the Bush administration*", using a paraphrase model trained on 490 million words of broadcast material, some of the generated paraphrase variants are:

Original sentence:
Economy is a big problem for the Bush administration
In-domain Paraphrases:
Economy is an uphill battle very much for the White House
Economy is a major issue for the American administration
Economy will be a main problem for the United States
Economy is very difficult indeed for the president
Economy is a real challenge for the administration
Economy is a significant problem for the Bush White House
Economy is another tremendous problem for the president
Economy is not good news for the administration
Economy represents a big trouble for the Bush presidency
Economy constitutes a large problem for the Bush government

As discussed in section 1, in order to improve the in-domain LM performance for conversational speech using the out-of-domain broadcast data, the domain independent and dependent characteristics of the two training data sources can be structurally exploited. Two cross-domain paraphrase generation methods are considered for this purpose:
1) Paraphrasing CTS data using BN paraphrase model: this approach allows the domain independent paraphrastic mappings that are learnt from the out-of-domain BN corpus, but not present in the in-domain data, to be used to generate an additional rich set of paraphrase variants for the conversational data, while retaining the same sentential structure and topic coverage. These variants are expected to further improve the in-domain context coverage when used in paraphrastic LM training. For the same example conversational sentence, using the BN data estimated paraphrase model, the following cross-domain paraphrase variants were found in the paraphrase lattice:

Cross-domain Paraphrases:
Actually I love
But largely I you know prefer
And I honestly generally hope
And quite frankly I probably prefer
Or I just really intend
And possibly I choose
Seemed like I always very much like
And I tend to probably prefer
And even more remarkably I usually prefer
And personally I generally want

2) Paraphrasing BN data using CTS paraphrase model: this approach allows the out-of-domain BN texts to be transformed into "in-domain like" data via a directed paraphrasing by restraining the choice of target paraphrases used to be found only in the in-domain data. Domain specific characteristics associated with the conversational data, such as disfluency and informal style, can be injected into the resulting paraphrase variants. These are expected to have a reduced domain mismatch against the in-domain data when used in paraphrastic LM estimation. For the example broadcast news sentence above, using the conversational data trained paraphrase model, the following cross-domain paraphrase variants were among those generated:

Cross-domain Paraphrases:
Economy is a heck of a uh problem for the president
Economy is a big big deal for the president
Economy is an awful big problem for I mean president
Economy it seems like a problem for uh Bush administration
Economy is like a big problem for uh the Bush administration
Economy that's like a uh problem for the president
Economy is you know a big problem for the president
Economy I I know is a big problem for the Bush administration
Economy 'cause I think a big problem for our president
Economy of course that's a big problem I think the president

The resulting paraphrase variants generated using the above two methods are then used to estimate cross-domain paraphrastic LM probabilities $P^{\mathsf{XD}}_{\mathsf{PLM}}(\cdot)$. These are then linearly combined with the baseline n-gram LM and standard PLM trained using in-domain paraphrases only. The interpolated LM probabilities in equation (3) is thus modified as,

$$P(\tilde{w}|\tilde{h}) = \lambda_{\mathsf{NG}} P_{\mathsf{NG}}(\tilde{w}|\tilde{h}) + \lambda_{\mathsf{PLM}} P_{\mathsf{PLM}}(\tilde{w}|\tilde{h}) + \lambda^{\mathsf{XD}}_{\mathsf{PLM}} P^{\mathsf{XD}}_{\mathsf{PLM}}(\tilde{w}|\tilde{h})$$

where $\lambda^{\mathsf{XD}}_{\mathsf{PLM}}$ is the interpolation weight assigned to the cross-domain paraphrastic LM. In common with equation (3), the interpolation weights can be optimized on held-out data.

5. Experiments and Results

In this section performance of cross-domain paraphrastic language models are evaluated on the CU-HTK LVCSR system for conversational telephone speech used in the 2004 DARPA EARS evaluation. The acoustic models were trained on approximately 2000 hours of Fisher conversational speech released by the LDC. A 59k recognition word list was used in decoding. The system uses a multi-pass recognition framework. In the initial lattice generation stage, adapted gender dependent cross-word triphone MPE acoustic models with HLDA projected, conversational side level normalized PLP features, and an interpolated 3-gram word level baseline LM were used. A detailed description of the baseline system can be found in [4]. The 3 hour **dev04** data, which includes 72 Fisher conversations, was used as a test set. For all results presented in this paper, matched pairs sentence-segment word error (MAPSSWE) based statistical significance test was performed at a significance level $\alpha = 0.05$.

The baseline LM was trained using a total of 1.0 billion words from 3 text sources combined at model level using perplexity optimized interpolation weights: the LDC Fisher acoustic transcriptions, **Fisher**, of 20 million words (0.75), and the University Washington conversational web data [3], **UWWeb** of 525 million words (0.18), and the out-of-domain broadcast news data **BN** of 490M words (0.07). The **BN** corpus include the PSM broadcast news transcripts from 1992 to 1999, broadcast news acoustic training transcripts from 1997 to 1998, Marketplace transcripts, TDT2, TDT3 and TDT4 closed captions from 2000 to 2001, the LDC broadcast news closed captions released in 2003 and CNN transcripts web collected from 1999 to 2003. The same three data sources were also used to build various standard and cross-domain paraphrastic language models. These LMs are then used for lattice rescoring and word error rate (WER) performance evaluation.

Information on corpus size, paraphrase extraction schemes used and the number of phrase pairs extracted from the these two text sources, as well as WordNet, are given in table 1. Using the automatic n-gram paraphrase extraction scheme presented in section 3, a total of 90k and 2.9M phrase pairs were extracted

from the in-domain **Fisher** and **UWWeb** data respectively. A total of 2.9M phrase pairs were also learnt from the out-of-domain **BN** data. The expert semantic labelling by WordNet, including synonyms, hypernyms, hyponyms and pertainyms, were used to generate 480k paraphrase phrase pairs.

Source	Size	In-domain	Extraction	# Phrase Pairs
WordNet	-	×	Expert	480k
Fisher	20M	√	Automatic	90k
UWWeb	525M	√	Automatic	2.9M
BN	490M	×	Automatic	2.9M

Table 1: Text size, domain description, paraphrase extraction method and the number of phrase pairs learnt from data sources.

The n-gram miss rates of three 4-gram word level LMs trained using the above three text sources are shown in table 2 on the reference transcription of **dev04**. The first LM is the baseline 3-way interpolated 4-gram LM without using any form of paraphrastic modelling. Using this baseline LM, the 3-gram and 4-gram miss rates are 17.9% and 49.4% respectively. When using a comparable paraphrastic 4-gram LM, trained using only in-domain paraphrases, but no cross-domain generated paraphrases as presented in section 4, the 3-gram and 4-gram miss rates were reduced by 13%-20% relative, to 14.3% and 42.9% respectively. At the same time the number of n-grams was increased by approximately a factor of four. As expected, if also using cross-domain paraphrases in training, the resulting cross-domain 4-gram paraphrastic LM, containing almost twice the number of n-grams, further reduced the 3-gram and 4-gram miss rates to 13.1% and 39.8%, by 19%-27% relative over the baseline 4-gram LM. These results suggest modelling cross-domain paraphrases can make more effective use of out-of-domain data to improve the in-domain data context coverage.

LM	Paraphrastic	Cross-domain	Miss Rate(%)	
			3g	4g
w4g	×	×	17.9	49.4
	√	×	14.3	42.9
	√	√	13.1	39.8

Table 2: n-gram miss rate of various LMs on **dev04**. "w4g" denotes a word level 4-gram LM.

WER performance of various LMs trained using the above three data sources are shown in table 3. The first three baseline LMs are non-paraphrastic. The word level 4-gram baseline LM "w4g" gave a WER of 16.6%. When further interpolated with a class based LM of 1000 automatically derived word clusters[11], the "w4g+clslm" model reduces the error rate by 0.2% absolute. The third baseline LM in table 3 is a multi-level LM, "w4g ∘ p4g", which incorporates phrase level linguistic constraints by log-linearly combining the word and phrase level 4-gram LMs. It was constructed by adding a total of 16k distinct multi-word phrases found in the **Fisher** data generated paraphrase phrase table to the baseline 59k word list, and trained on the phrase level segmented text data. This is similar to the method used in [21]. Word level lattices need to be first converted to phrase level lattices when using the multi-level LM. This was implemented using a WFST composition between the word level lattice with the phrase level segmentation transducer. After the log-linear combination between word and phrase level LMs is performed, the resulting phrase level lattices are converted back to word level again via a WFST composition with the phrase to word transducer, to obtain the 1-best word level hypothesis for WER evaluation. By adding additional phrase level features, this multi-level LM gives a WER reduction of 0.2% absolute over the word level 4-gram baseline LM.

In the second section of table 3, WER performance of the standard word and multi-level paraphrastic LMs, constructed only in-domain, but no cross-domain paraphrasing, are shown in the 4th and 5th lines respectively. Compared with their non-paraphrastic baselines shown in the 1st and 3rd line of table 3, a consistent WER reduction of 0.3% was obtained in both cases. If also using cross-domain generated paraphrases in training, further WER reductions were obtained for both the word and multi-level cross-domain paraphrastic LMs, as shown in the last two lines of table 3. The lowest WER was produced by the multi-level cross-domain paraphrastic LM, shown in the 7th line of table 3. An overall statistically significant WER reduction of 0.6% absolute was obtaind over the word level 4-gram baseline LM constructed using standard model interpolation.

LM	Paraphrastic	Cross-domain	dev04
w4g	×	×	16.6
w4g+clslm			16.4
w4g ∘ p4g			16.4
w4g	√	×	16.3
w4g ∘ p4g			16.1
w4g	√	√	16.2
w4g ∘ p4g			16.0

Table 3: Performance of LMs trained on **dev04**. "w4g" denotes a word level 4-gram LM. "w4g+clslm" a word level 4-gram LM interpolated with a class LM with 1000 classes, and "w4g ∘ p4g" a multi-level LM log-linearly combining word and phrase level 4-gram LMs. Other naming convention same as table 2.

6. Conclusion

Cross-domain paraphrastic language models were investigated in this paper to improve in-domain language model performance using out-of-domain data. Significant error rate reduction of 0.6% absolute were obtained on a state-of-the-art large vocabulary conversational speech recognition task. Experimental results suggest the proposed method can improve the in-domain LM context coverage and generalization and thus may be useful for speech recognition in under-resourced domains. Future research will focus on improving paraphrase pair extraction, modelling method and cross-domain paraphrasing.

7. References

[1] I. Androutsopoulos & P. Malakasiotis (2010). A Survey of Paraphrasing and Textual Entailment Methods, *Journal of Artificial Intelligence Research*, 38:135-187, 2010.

[2] R. Barzilay & L. Lee (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment, in *Proc. of HLT-NAACL 2003*, pp 16-23, Edmonton.

[3] I. Bulyko, M. Ostendorf & A. Stolcke. Getting More Mileage from Web Text Sources for Conversational Speech Language

Modeling using Class-Dependent Mixtures, in *Proc. HLT'03*, Edmonton.

[4] G. Evermann et al. (2004). Training LVCSR Systems on Thousands of Hours of Data, in *Proc. ICASSP2005*, Philadelphia.

[5] C. Fellbaum (1998) *WordNet: An Electronic Lexical Database*, MIT Press. Cambridge, MA.

[6] Z. Harris (1954). Distributional Structure, *Word*, 10(2):3 pp.146-162.

[7] J. B. Hsu (2009). *Language Modeling for Limited-data Domains*, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

[8] R. Iyer & M. Ostendorf (1997). Transforming Out-of-domain Estimates to Improve In-domain Language Models, in *Proc. Eurospeech97'*, vol. 4, pp. 1975-1978, Rhodes, Greece.

[9] R. Iyer, M. Ostendorf & H. Gish (1997) Using Out-of-domain Data to Improve In-domain Language Models, in *IEEE Signal Processing Letters*, vol.4, no.8, pp. 221-223, August 1997.

[10] R. Iyer & M. Ostendorf (1999). Relevance Weighting for Combining Multi-domain Data for n-gram Language Modeling, *Computer Speech & Language*, Volume 13, Issue 3, pp. 267-282, 1999.

[11] R. Kneser & H. Ney (1993), "Improved clustering techniques for class based statistical language modeling," in *Proc. EuroSpeech93'*, Berlin.

[12] D. Lin & P. Pantel (2001). DIRT - Discovery of Inference Rules from Text, in *Proc. ACM SIGKDD2001*, pp.323-328, San Francisco, CA.

[13] X. Liu, J. L. Hieronymus, M. J. F. Gales & P. C. Woodland (2010). Language Model Combination and Adaptation Using Weighted Finite State Transducers, in *Proc. IEEE ICASSP2010*, pp. 5390-5393, Dallas.

[14] X. Liu, M. J. F. Gales & P. C. Woodland (2013). Use of Contexts in Language Model Interpolation and Adaptation, *Computer Speech and Language*, Volume 27, Issue 1, pp. 301-321, January 2013.

[15] X. Liu, J. L. Hieronymus, M. J. F. Gales & P. C. Woodland (2013). Syllable Language Models for Mandarin Speech Recognition: Exploiting Character Sequence Models, *Journal of the Acoustical Society of America*, Volume 133, Issue 1, pp. 519-528, January 2013.

[16] X. Liu, M. J. F. Gales & P. C. Woodland (2012). Paraphrastic Language Models, in *Proc. ISCA Interspeech2012*, Portland, Oregon.

[17] X. Liu, M. J. F. Gales & P. C. Woodland (2013). Paraphrastic Language Models and Combination with Neural Network Language Models, to appear in *Proc. IEEE ICASSP2013*, Vancouver, Canada.

[18] N. Madnani & B. Dorr (2010). Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods, *Computational Linguistics*, Vol. 36, No. 3, 2010.

[19] M. Mohri. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23:2, 1997.

[20] T. R. Niesler , E. W. D. Whittaker & P. C. Woodland (1998) Comparison Of Part-Of-Speech And Automatically Derived Category-Based Language Models For Speech Recognition, in *Proc. IEEE ICASSP1998*, Vol.1, pp. 177-180, Seattle, WA.

[21] M. Padmanabhan et al. (1998). Speech Recognition Performance on a Voicemail Transcription Task, In *Proc. IEEE ICASSP1998*, Vol. 2 pp. 913-916, Seattle, WA.

[22] M. Pasca & P. Dienes (2005). Aligning needles in a haystack: Paraphrase acquisition across the Web, In *Proc. IJCNLP2005*, pp. 119-130, Jeju Island.

[23] A. Rudnicky (1995). Language Modeling with Limited Domain Data, in *Proc. 1995 ARPA Workshop on Spoken Language Technology*, pp. 66-60, Morgan Kaufmann.

[24] F. Weng, A. Stolcke & A. Sankar (1997). Hub4 Language Modeling Using Domain Interpolation and Data Clustering, in *Proc. 1997 DARPA Speech Recognition Workshop*, pp. 147-151, Westfields International Conference Center, Chantilly, Virginia.

[25] P. C. Woodland, M. J. F. Gales, D. Pye & S. J. Young (1996). The development of the 1996 HTK broadcast news transcription system, in *Proc. 1996 DARPA Speech Recognition Workshop*, pp. 73-78, Arden House, New York.